Bezpieczeństwo cybernetyczne

W trakcie czytania ustaliliśmy jedno. Aby zrealizować sztuczną inteligencję, potrzebujemy dostępu do dużych ilości danych. Dane odgrywają kluczową rolę w budowaniu możliwości inteligentnych maszyn, które uzupełniają i zwiększają ludzkie możliwości. Aplikacje, które tworzymy w oparciu o architektury i algorytmy uczenia maszynowego są tak dobre, jak dane bazowe. Wraz ze wzrostem naszej zależności od danych zaczynamy postrzegać dane jako zasób w przypadku systemów o kluczowym znaczeniu, takich jak sprzęt medyczny, lotnictwo, systemy bankowe itp., zachowanie integralności zasobów danych jest jednym z najważniejszych priorytetów i kluczowych składników udanego rozpowszechnienia systemów opartych na sztucznej inteligencji. Ochrona infrastruktury krytyczna związana z naruszeniami danych jest ogólnie znana jako bezpieczeństwo cybernetyczne. Tu zobaczymy, w jaki sposób możemy wykorzystać różne ramy zarządzania danymi, aby chronić krytyczne zasoby danych i wykorzystać nasze zrozumienie ram zarządzania Big Data i uczenia maszynowego, aby zabezpieczyć nasz najważniejszy zasób (dane). Zajmiemy się następującymi tematami:

* Jak możemy wykorzystać Big Data do ochrony infrastruktury krytycznej

* Ogólne koncepcje przetwarzania strumieniowego

* Informacje o bezpieczeństwie i zarządzanie zdarzeniami

* Struktura pliku dziennika dostępu do serwera sieci Web i strategie wykorzystania go do bezpieczeństwa cybernetycznego

* Splunk jako aplikacja korporacyjna do wdrażania bezpieczeństwa cybernetycznego

* ArcSight jako platforma zarządzania bezpieczeństwem przedsiębiorstwa

Big Data do ochrony infrastruktury krytycznej

Infrastruktura krytyczna (CI) to termin używany przez przedsiębiorstwa i agencje rządowe do definiowania aktywów i modeli roboczych, które muszą funkcjonować na optymalnym poziomie, aby zapewnić płynne i harmonijne wrażenia dla interesariuszy, którzy bezpośrednio lub pośrednio korzystają lub są pod wpływem te systemy. Przykłady obejmują sieć energetyczną, zaopatrzenie w wodę, transport, organy ścigania i wiele takich systemów, które muszą działać bezproblemowo przez całą dobę. W ciągu ostatnich kilku dziesięcioleci większość IK uległa digitalizacji i generuje coraz więcej danych z heterogenicznych źródeł. Te dodatkowe zasoby danych skutkują ciągłą poprawą i eliminują potrzebę interwencji człowieka, a tym samym zmniejszają błąd. Dane generowane przez te systemy są wykorzystywane jako zasób analizy opisowej i predykcyjnej w celu planowania konserwacji zapobiegawczej i zapobiegania awariom. Dzięki podejściu opartemu na danych do podstawowego funkcjonowania CI, zaobserwowaliśmy ogromną poprawę wydajności i ogólnej niezawodności CI. Jednak zdarzają się ogromne incydenty, w których atakujący o złośliwych zamiarach zakłócenia CI z powodzeniem włamali się do CI i spowodowali zakłócenia. Na przykład Stuxnet, który został znaleziony w 2010 roku, celował w systemy SCADA (nadzór nadzorczy i pozyskiwanie danych) i spowodował uszkodzenie planów wzbogacania paliwa w Iranie, współpracując z programowalnymi sterownikami logicznymi (PLC). Istnieje wiele takich incydentów i prób, które zakłócają CI i powodują wieczyste szkody. Jednym z najważniejszych aspektów zapobiegania atakom cyberbezpieczeństwa na CI jest dostępność danych z CI, które są generowane w środowisku pracy. Dane te muszą być dostępne do analizy i potencjalnych działań możliwie jak najbliżej zdarzenia. Wraz z danymi z głównych komponentów CI, dane z innych heterogenicznych systemów, które są pośrednio powiązane z CI, muszą zostać wykorzystane do zbudowania solidnego mechanizmu obrony przed cyberatakami. Oznacza to, że potrzebujemy objętości danych, prędkości i różnorodności, aby skutecznie chronić CI.

Te trzy V wraz z wartością jako czwarte V, która jest uzyskiwana z danych razem, stanowią Big Data. Innymi słowy, Big Data jest kluczowym zasobem dla skutecznych strategii przeciw cyberatakom. Potrzebujemy stale ewoluujących ram i procesów opartych na danych do ochrony CI, wykorzystując analitykę Big Data do skutecznego monitorowania bezpieczeństwa i ochrony. Ta struktura oparta na danych ma trzy główne elementy, jak pokazano na poniższym diagramie:

Gromadzenie i analiza danych

Podstawowe systemy tworzące elementy CI generują zasoby danych w postaci dzienników zdarzeń. Komponent do gromadzenia danych musi gromadzić te dzienniki ze wszystkich komponentów (oprogramowania i sprzętu). Oprócz systemów podstawowych proces powinien również gromadzić dane ze środowiska kontekstowego systemów CI. Niejednorodne logi pomagają w holistycznej analizie oraz dokładniejszej i dokładniejszej rozdzielczości. Oprócz uruchomionych dzienników system powinien także mieć możliwość przechowywania danych historycznych dla systemów CI i uzyskiwania do nich dostępu. Dane historyczne zapewniają wgląd oparty na podobieństwie wzorców do przeszłych zdarzeń. Jeśli wcześniejsze ograniczenia doprowadziły do szybkiej korekty i rozwiązania krytycznego zdarzenia, można zastosować nadzorowane uczenie się w celu podjęcia podobnych działań w oparciu o doświadczenie. Dane historyczne również bardzo pomagają w zapobieganiu przyszłym atakom opartym na podobnych lukach w systemie. Dane (log) generowane przez komponenty CI i powiązany kontekst środowiskowy można podzielić na trzy typy:

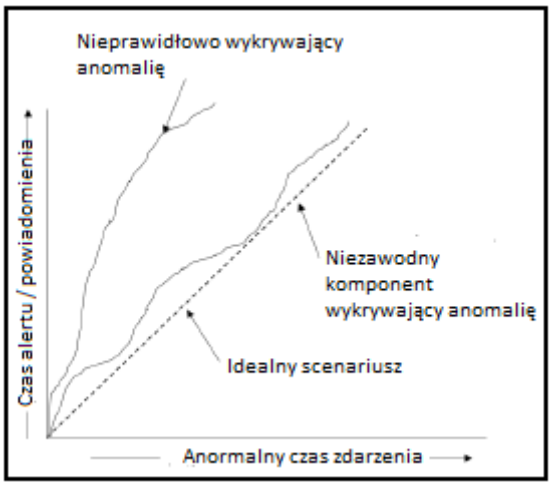
* Dane strukturalne: W przypadku formatu strukturalnego poszczególne elementy (atrybuty) bytu są reprezentowane w z góry określony i spójny sposób w różnych okresach. Na przykład dzienniki generowane przez serwery WWW (dziennik HTTP) reprezentują pola takie jak adres IP, czas, w którym serwer zakończył przetwarzanie żądania, metodę HTTP, kod stanu itd. Wszystkie te atrybuty żądania internetowego są konsekwentnie reprezentowane w żądaniach. Ustrukturyzowane dane są stosunkowo łatwe do przetworzenia i nie wymagają skomplikowanego analizowania i przetwarzania wstępnego, zanim będą dostępne do analizy. Dzięki uporządkowanym danym przetwarzanie jest szybkie i wydajne.

* Dane nieustrukturyzowane: jest to swobodnie płynący format dziennika aplikacji, który nie jest zgodny z żadnymi predefiniowanymi regułami strukturalnymi. Te dzienniki są zwykle generowane przez aplikacje i mają być używane przez osobę, która rozwiązuje problemy. Intencją jest rejestrowanie zdarzeń bez wyraźnego celu, aby dzienniki były czytelne dla komputera. Te dzienniki wymagają obszernego przetwarzania wstępnego, analizy i pewnej formy przetwarzania języka naturalnego, zanim będą one dostępne do analizy.

* Dane częściowo ustrukturyzowane: Jest to kombinacja danych ustrukturyzowanych i nieustrukturyzowanych, w których niektóre atrybuty w formacie ustrukturyzowanym są reprezentowane w nieustrukturyzowany sposób. Informacje są podzielone na pola, które można łatwo przeanalizować, ale poszczególne pola wymagają dodatkowego przetwarzania wstępnego, zanim zostaną wykorzystane w analizie.

Wykrywanie anomalii

Gdy zaczynamy zbierać dane z systemów heterogenicznych, istnieje pewien wzorzec, który jest ustalany pod względem wielkości danych, struktury, zawartości informacji i prędkości danych. Wzór ten pozostaje spójny w standardowych warunkach pracy i można spodziewać się skoków lub zmian wzorów. Na przykład sprzedawca online może spodziewać się większej liczby zamówień w okresie świątecznym, a to wydarzenie nie jest traktowane jako anomalia. W przypadku nieoczekiwanej zmiany w regularnym schemacie danych pod względem objętości, prędkości i różnorodności element wykrywający anomalię uruchamia alert i powiadomienie. Jedną z ważnych cech znacznie rozwiniętego i niezawodnego komponentu wykrywającego anomalię jest to, że jest on w stanie wygenerować alert natychmiast po wystąpieniu zdarzenia, z minimalnym opóźnieniem między czasem zdarzenia a czasem alarmu / powiadomienia. Poniższy schemat przedstawia idealne, niezawodne i niewiarygodne wykrywanie anomalii składniki oparte na różnicy czasu między zdarzeniem a czasem alarmu:

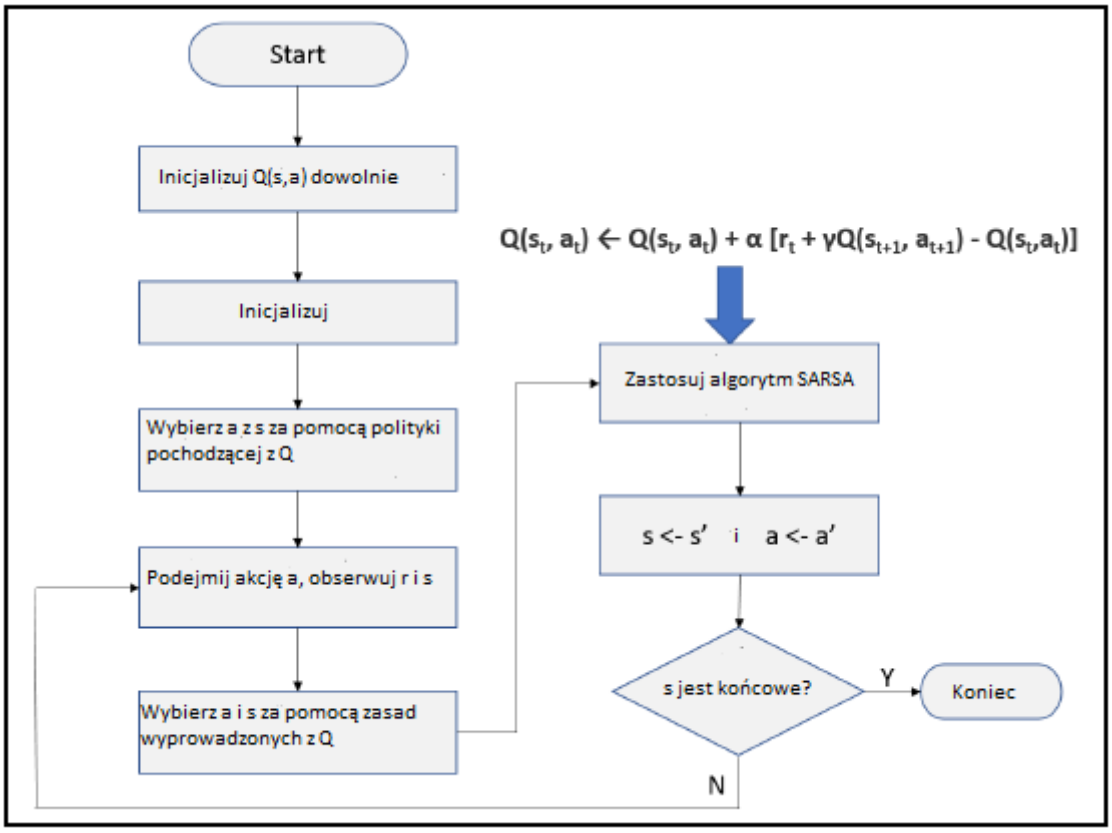


Działania korygujące i zapobiegawcze

Gdy podejrzana aktywność zostanie wykryta przez komponent wykrywający anomalię, istnieją dwa sposoby zareagowania. W pierwszym przypadku alert / powiadomienie wymaga ręcznej interwencji w celu uruchomienia działania naprawczego. W drugim przypadku sam system podejmuje pewne działania naprawcze w oparciu o kontekst i dopuszczalny próg marginesu błędu. Na przykład, jeśli włamanie do obwodu termostatu zacznie nieprzewidziany wzrost temperatury w chłodni, system może przełączyć sterowanie na alternatywny termostat i upewnić się, że temperatura wróciła do normy i utrzymywała się na normalnym poziomie. Ten komponent może wykorzystywać nadzorowane uczenie się, a także algorytmy uczenia wzmacniającego do samodzielnego uruchamiania działań naprawczych na podstawie danych historycznych lub funkcji nagrody. Po zastosowaniu korekty i przywróceniu stanu CI do normalnego stanu system musi przeanalizować główną przyczynę i sam się szkolić, aby podejmował działania zapobiegawcze (zastosowanie poprawki, zmiany w modelu bezpieczeństwa, wdrożenie nowych kontroli dostępu itd.) .

Koncepcyjny przepływ danych

W typowych środowiskach Big Data implementowana jest architektura warstwowa. Warstwy w potoku przetwarzania danych pomagają oddzielić poszczególne etapy, przez które przechodzą dane, aby chronić krytyczną infrastrukturę. Dane przepływają przez przyjmowanie, przechowywanie, przetwarzanie i cykl akcji, co przedstawiono na poniższym rysunku wraz z popularnymi strukturami używanymi do wdrażania przepływu pracy:



Większość komponentów użytych na tym rysunku to oprogramowanie typu open source i jest wynikiem współpracy dużej społeczności. Szczegółowa dyskusja na temat wszystkich tych elementów nie wchodzi w zakres tego rozdziału. Pozwól nam jednak zrozumieć te elementy na wysokim poziomie w kontekście bezpieczeństwa cybernetycznego.

Przegląd komponentów

Aby pomyślnie wdrożyć strategię ochrony CI, konieczne jest zebranie danych z heterogenicznych źródeł poza oczywistymi źródłami, takimi jak logi serwera. Im więcej źródeł danych zostanie zidentyfikowanych i zintegrowanych, tym większa jest potrzeba przechowywania. Biorąc pod uwagę wielkość i prędkość danych, nie jest możliwe dostosowanie danych przy użyciu tradycyjnych systemów plików. Zamiast tego nowoczesna architektura wykorzystuje rozproszone systemy plików.

Rozproszony system plików Hadoop

Rozproszony system plików Hadoop (HDFS) jest jedną z najpopularniejszych implementacji rozproszonego systemu plików. Jest to rdzeń Hadoop, która jest platformą przetwarzania rozproszonego. System HDFS został zaprojektowany i ewoluował z myślą o następujących celach, które uzupełniają wymagania dotyczące pamięci masowej w celu ochrony CI:

* Awaria sprzętowa: HDFS replikuje każdy blok pliku na trzech (domyślnych) węzłach. Podstawową ideą korzystania z przetwarzania rozproszonego jest możliwość wykorzystania sprzętu, a zatem klaster

składa się z dużej liczby stosunkowo niewielkich węzłów. Przy dużej liczbie węzłów prawdopodobieństwo awarii n de wzrasta. Jednym z nich jest wykrywanie i usuwanie awarii sprzętu bez utraty danych głównych celów HDFS. Systemy ochrony CI potrzebują również tego samego poziomu niezawodności i odporności na uszkodzenia w celu wykrycia zagrożeń bezpieczeństwa cybernetycznego.

* Duże zestawy danych: zakłada się, że aplikacje wykorzystujące HDFS jako bazowe magazyny danych zajmują się dużymi zestawami danych w zakresie od wielu gigabajtów do terabajtów i więcej. HDFS jest wbudowany w obsługę dużych plików danych. Systemy ochrony CI generują również duże ilości danych i radzą sobie z nimi. Dobrym przykładem jest centralna władza kraju, który monitoruje sieć internetową tego kraju i zajmuje się setkami gigabajtów danych na sekundę.

* Prosty model koherencji: aplikacje CI generują pliki dziennika, które należy zapisać raz i odczytać wiele razy. Model koherencji jest także jednym z głównych celów projektowania HDFS. Plik, raz utworzony i zapisany, nie musi być zmieniany w tym modelu. Cel ten uzupełnia również aplikacje bezpieczeństwa cybernetycznego.

* Przenośność na heterogenicznych platformach sprzętowych i programowych: HDFS można łatwo przenosić na różne platformy. Cel ten uzupełnia również podstawowe wymaganie systemów bezpieczeństwa cybernetycznego. Systemy bezpieczeństwa cybernetycznego są wdrażane na różnych platformach, a przenośność systemu plików HDFS jako podstawowego systemu plików może być dodatkową zaletą.

Bazy danych NoSQL

NoSQL (nie tylko SQL) to paradygmat, w którym dane są przechowywane w postaci encji zamiast typowego tabelarycznego formatu relacyjnego typu RDBMS. Jednym z głównych celów baz danych NoSQL jest skalowanie w poziomie i wysoka dostępność. W oparciu o bazową strukturę danych baz danych NoSQL są one podzielone na:

* Bazy danych dokumentów: Każdy klucz w bazie danych jest mapowany na dokument. Dokument może być plikiem binarnym lub zagnieżdżoną strukturą, taką jak XML lub JSON. Przykładami baz danych dokumentów są MongoDB, CouchDB, Couchbase i tak dalej.

* Bazy danych wykresów: Przydają się w przypadku danych w postaci połączonych wykresów, takich jak połączenia z mediami społecznościowymi. Przykładami baz danych grafów są Neo4j, OrientDB, Apache Giraph i tak dalej.

* Kolumnowe bazy danych: te bazy danych reprezentują dane, przechowując dane kolumn razem zamiast wierszy. Są zoptymalizowane do przechowywania rozproszonego i szybkiego dostępu do zapytań w bardzo dużych bazach danych. Przykładami kolumnowych baz danych są Cassandra, HBase i tak dalej.

Bazy danych NoSQL mogą być skutecznie wykorzystywane w implementacjach aplikacji bezpieczeństwa cybernetycznego, ponieważ mogą łatwo obsługiwać duże ilości struktur oraz częściowo ustrukturyzowanych i nieustrukturyzowanych danych, które są gromadzone z heterogenicznych źródeł otaczających CI. Bazy danych NoSQL obsługują również architekturę rozproszoną geograficznie, którą można skalować na żądanie bez wpływu na już utrwalone dane. Ta funkcja jest przydatna w przypadku stopniowego wzrostu infrastruktury CI, takiej jak usługi telekomunikacyjne w odległych obszarach, które są stopniowo budowane.

MapReduce

MapReduce (MR) to paradygmat programowania stanowiący rdzeń Hadoop. Może skalować przetwarzanie danych do ogromnie dużych woluminów. Dane i przetwarzanie mogą być dystrybuowane do setek i tysięcy węzłów w celu skalowania poziomego. Jak sama nazwa wskazuje, zadania MR zawierają dwie fazy:

* Faza mapy

* Faza redukcji

W fazie mapy zestaw danych jest dzielony na części i wysyłany do niezależnego procesu w celu zebrania wyniku. Te równoległe procesy mapujące działają niezależnie na różnych dostępnych węzłach w klastrze. Po zakończeniu przetwarzania (zadanie mapy) wyniki są tasowane i sortowane przed rozpoczęciem zadań zmniejszania. Zadania zmniejszania ponownie uruchamiają się niezależnie w dostępnych węzłach, a całe obliczenia są wykonywane jako całość. Wyniki pośrednie są przechowywane w systemie plików (HDFS) i obejmują operacje we / wy. Z powodu tych operacji we / wy paradygmat MR jest odpowiedni dla obciążeń zorientowanych wsadowo, w których mają być przetwarzane bardzo duże ilości zestawów danych. W kontekście bezpieczeństwa cybernetycznego szkielet MR może być wykorzystywany do przetwarzania danych historycznych pochodzących z CI oraz otaczającego kontekstu aplikacji i środowiska. Dane mogą być agregowane do raportowania i mogą być wykorzystane jako dane szkoleniowe do nadzorowanego wdrożenia cyberbezpieczeństwa opartego na uczeniu się.

Apache Pig

HDFS i MR są silnikami pamięci i obliczeniowymi stanowiącymi rdzeń Hadoop. Surowa implementacja aplikacji przetwarzania równoległego jest złożona i podatna na błędy. Apache Pig zapewnia owijanie wokół zadań przetwarzania równoległego w Hadoop. Pig ułatwia przetwarzanie dużych zestawów danych, zapewniając prosty interfejs programowania i interfejs API. Zadania i działania napisane za pomocą Pig są z natury równoległe do klastra Hadoop. W kontekście bezpieczeństwa cybernetycznego Pig może być wykorzystywany do realizacji złożonych równoległych zadań agregacji danych i wykrywania anomalii wraz z przygotowaniem danych szkoleniowych do nadzorowanego uczenia się w przypadku, gdy aplikacja ochrony CI wykorzystuje algorytmy uczenia maszynowego.

Hive

Apache Hive to hurtownia danych zbudowana na platformie Hadoop. Hive zapewnia interfejs podobny do SQL dla danych rezydujących na HDFS. Zapytania są wykonywane jako zadania MR, Tez lub Spark w klastrze Hadoop. Hive obsługuje indeksowanie szybkich zapytań wraz ze skompresowanymi typami pamięci, takimi jak ORC. W kontekście cyberbezpieczeństwa Hive może służyć do przechowywania zagregowanych widoków różnych dzienników generowanych przez aplikacje CI. Podczas gdy struktury przetwarzania wsadowego, takie jak MR na platformie Hadoop, są przydatne w efektywnym przetwarzaniu bardzo dużych ilości danych, nie są one odpowiednie do zapewnienia bezpieczeństwa CI elementów misji. Takie systemy CI wymagają przetwarzania w czasie rzeczywistym (przynajmniej prawie w czasie rzeczywistym) danych przesyłanych strumieniowo lub mikroprocesorów w celu uzyskania szybkich alertów, powiadomień i działań na czas. Architektura przetwarzania strumieniowego wymaga większej koncentracji w kontekście bezpieczeństwa cybernetycznego i ochrony CI.

Opis przetwarzania strumieniowego

Aplikacje wdrażane w przedsiębiorstwie mają dwa podstawowe komponenty:

* Infrastruktura

* Aplikacje

Infrastruktura obejmuje fizyczny sprzęt i sieć, która łączy ze sobą różne systemy. Wdrożenie bezpieczeństwa infrastruktury i aplikacji ma różne uwarunkowania, z powodu których ramy i procesy ochrony CI są również różne. Systemy bezpieczeństwa muszą działać na peryferiach infrastruktury i wewnątrz aplikacji. Istnieją różne zdarzenia, przez które przepływają dane (sieć i aplikacja). Zdarzenia mają miejsce w określonym czasie, a odpowiednie dane są dostępne do analizy i działania natychmiast po wystąpieniu zdarzenia. Na przykład aplikacja kliencka, taka jak przeglądarka internetowa, żąda dostępu do strony internetowej za pośrednictwem protokołu HTTP. Sekwencja zdarzeń jest inicjowana bezpośrednio po wprowadzeniu adresu URL przez przeglądarkę. Powiązana analiza oparta na żądaniu musi odbywać się jak najbliżej czasu zdarzenia, aby chronić aplikację internetową przed złośliwymi atakami. Zdolność przetwarzania danych jako strumienia wykrywającego anomalie jest kluczowym czynnikiem do skutecznego wdrożenia bezpieczeństwa cybernetycznego. Kluczowymi zagadnieniami do przetwarzania strumieniowego są dane nieograniczone, przetwarzanie danych nieograniczonych i analiza oparta na niskim opóźnieniu:

* Dane nieograniczone: termin ten dotyczy praktycznie nieograniczonej liczby zestawów danych. Na przykład pakiety sieciowe, które przepływają z jednego systemu fizycznego do drugiego. Te pakiety zawierają informacje, które ciągle są generowane jako ciągły strumień.

* Nieograniczone przetwarzanie danych: przetwarzanie musi nastąpić, gdy dane są w ruchu. Pakiety sieciowe lub dane aplikacji muszą być dostępne i przetwarzane podczas ich generowania, w przeciwieństwie do silnika przetwarzania wsadowego, w którym dane trafiają do trwałej pamięci przed przetworzeniem.

*Analiza małych opóźnień: analiza oparta na danych niezwiązanych musi odbywać się możliwie jak najbliżej zdarzenia w przypadku przypadków użycia przesyłania strumieniowego. Cyberbezpieczeństwo to krytyczny przypadek użycia, który wymaga analizy o niskim opóźnieniu i działań, aby był skuteczny. Jak widzieliśmy na ryc. 11.2, wykrywanie anomalii jest niezawodne, gdy czas zdarzenia i czas alarmu / powiadomienia są oddzielone minimalnym przekrzywieniem. Ta różnica jest zmienna i zależy od wielu warunków, takich jak przeciążenie sieci, opóźnienie wprowadzone do narzutu przetwarzania w środowisku rozproszonym i tak dalej.

Semantyka przetwarzania strumieniowego

Gdy zdarzenia są wyzwalane w systemie, pojawiają się komunikaty (pakiety danych), które są generowane u źródła i przetwarzane w silnikach przetwarzania. Istnieją trzy różne semantyki dla systemów przetwarzania strumienia, co najmniej raz, co najwyżej raz i dokładnie raz:

* Co najmniej raz: w tym przypadku wiadomość może zostać wysłana przez źródło więcej niż raz. Jednak silnik przetwarzania musi zagwarantować, że jeden komunikat zostanie przetworzony co najmniej raz z wielu transmisji tego samego komunikatu. Możliwe jest, że wiadomość jest przetwarzana więcej niż raz i może być akceptowalna w niektórych przypadkach użycia. Aplikacja końcowa może wymagać uruchomienia sprawdzenia usuwania duplikatów semantyki.

* Co najwyżej raz: aplikacja do przetwarzania strumieniowego gwarantuje, że wiadomość zostanie przetworzona tylko raz. Nawet jeśli istnieje wiele transmisji tego samego komunikatu, silnik przetwarzania musi zagwarantować, że komunikat nie zostanie przetworzony więcej niż jeden raz. Może się zdarzyć, że dany pakiet w ogóle nie zostanie przetworzony, ale nie można go przetworzyć więcej niż raz. Ten semantyczny jest krytyczny w aplikacjach, w których wynik końcowy transakcji

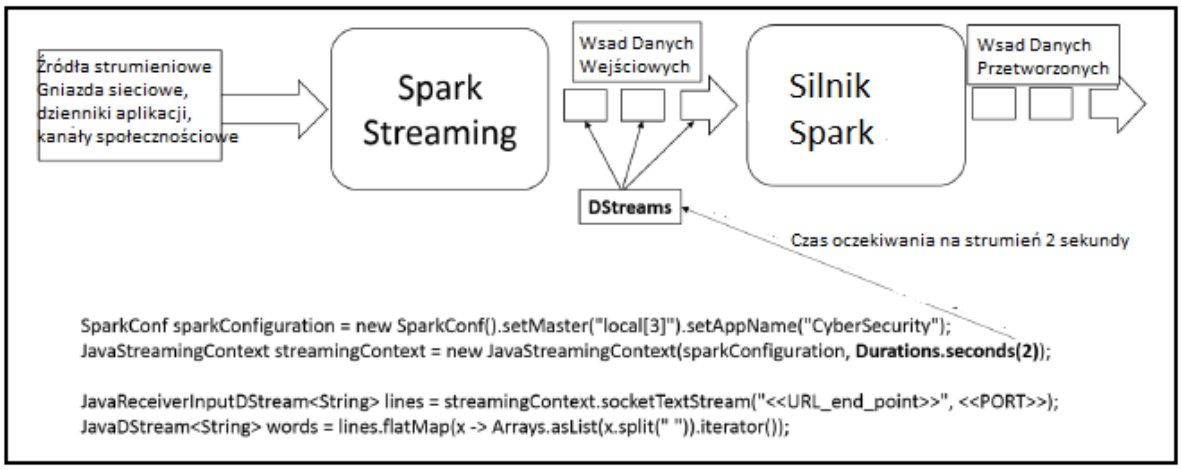
prowadzi do niespójnego stanu, jeśli komunikat jest przetwarzany więcej niż jeden raz. Na przykład transakcja bankowa z przekazem środków musi być ściśle przestrzegana co najwyżej raz semantycznie.

* Dokładnie raz: nawet jeśli system źródłowy dostarcza wiadomość więcej niż jeden raz, jest ona zużywana i przetwarzana dokładnie raz. Jest to najbardziej idealny semantyczny system cyberbezpieczeństwa. Krytyczny komunikat przetworzony tylko raz gwarantuje terminowe i właściwe działanie, które może zapobiec potencjalnym atakom na sieć i infrastrukturę aplikacji. Jednak ten semantyczny dokładnie raz jest najważniejszy i trudny do wdrożenia, ponieważ wymaga ścisłej współpracy między systemem źródłowym a docelowym. Silna spójność jest podstawowym wymogiem dla dokładnie raz semantycznej.

Dokładnie raz semantyczny proces przetwarzania danych strumieniowych jest obsługiwany przez niektóre platformy open source, takie jak Spark Streaming, Apache Kafka i Apache Storm. Pozwól nam zrozumieć te frameworki na wysokim poziomie, zanim spojrzymy na architekturę wysokiego poziomu systemu bezpieczeństwa cybernetycznego, która wykorzystuje te frameworki.

Spark Streaming

Spark jest silnikiem obliczeniowym rozproszonym ogólnego zastosowania w pamięci. Spark Streaming API to rozszerzenie podstawowej biblioteki Spark, która została zaprojektowana z myślą o skalowalności, wysokiej przepustowości i odporności na uszkodzenia w celu przesyłania strumieniowych (nieograniczonych) celów danych. Spark Streaming integruje się z różnymi źródłami danych, takimi jak gniazda sieciowe TCP, logi serwera HTTP, producenci kafka, strumienie mediów społecznościowych i tak dalej. Strumienie i złożone zdarzenia są przetwarzane za pomocą operacji ogólnych, takich jak MapReduce, join i okienkowanie. Dane w ruchu można analizować, agregować, filtrować i wysyłać do dalszych aplikacji, pamięci trwałej lub paneli kontrolnych na żywo. Algorytmy uczenia maszynowego i przetwarzania grafów oraz interfejsy API można zastosować do nieograniczonych danych za pomocą Spark Streaming. Spark Streaming dzieli dane strumieniowe na partie na podstawie okienkowania opartego na czasie. Strumień jest dzielony na fragmenty w określonych (wstępnie zdefiniowanych i konfigurowalnych) odstępach czasu i przetwarzany jako dyskretny strumień jako abstrakcyjna jednostka przetwarzająca na niskim poziomie. Nazywa się to DStream. DStreams można utworzyć na podstawie wejściowych danych strumieniowych (dzienniki sieciowe lub dzienników aplikacji) lub można je wykorzystać z systemów strumieniowych takich jak Flume, Storm lub Kafka. Strumień przesyłania strumieniowego Spark można postrzegać koncepcyjnie w następujący sposób:



Spark Streaming zapewnia dokładnie raz semantykę danych przesyłanych strumieniowo jako niezawodny odbiornik, gdy źródło przesyłania strumieniowego jest włączone do przetwarzania potwierdzeń (na przykład Kafka).

Kafka

Kafka działa jako dziennik z wyprzedzeniem zapisu, który rejestruje wiadomości w trwałym sklepie i pozwala subskrybentom czytać i stosować te zmiany we własnych sklepach w odpowiednim czasie systemowym. Do powszechnych subskrybentów należą usługi na żywo, które wykonują agregację komunikatów lub inne przetwarzanie tych strumieni, a także potoki Hadoop i hurtowni danych, które ładują praktycznie wszystkie kanały do przetwarzania zorientowanego na partię. Ogólnie rzecz biorąc, Kafka został zbudowany z następujących elementów mając na uwadze:

* Luźne powiązanie między producentami wiadomości a konsumentami wiadomości

* Trwałość danych wiadomości dla różnych konsumentów i obsługa awarii

* Maksymalizuj przepustowość od końca do końca dzięki komponentom o niskim opóźnieniu

* Zarządzanie różnorodnymi formatami i typami danych

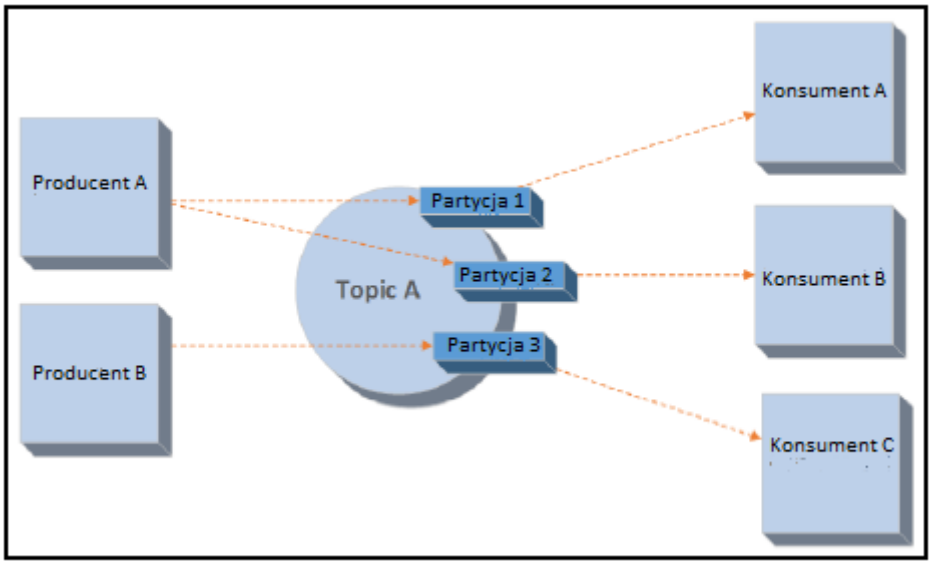
* Skalowanie serwerów liniowo bez wpływu na istniejącą konfigurację

W Kafce każda wiadomość jest tablicą bajtów. Producenci to aplikacje lub procesy, które chcą przechowywać informacje w kolejkach Kafka. Wysyłają wiadomości do tematów Kafka, które przechowują wiadomości wszystkich typów. Każdy temat jest podzielony na jedną lub więcej partycji. Każda partycja jest uporządkowanym dziennikiem komunikatów z zapisem z wyprzedzeniem. System wykonuje tylko dwie operacje:

* Aby dołączyć na końcu dziennika

* Aby pobrać wiadomości z danej partycji, zaczynając od identyfikatora wiadomości

Fizycznie każdy temat jest rozłożony na różnych brokerów Kafka, którzy obsługują jedną lub dwie partycje każdego tematu. Idealnie, rurociągi Kafka powinny mieć jednolitą liczbę partycji na brokera i wszystkie tematy na każdym komputerze. Konsumenci to aplikacje lub procesy, które subskrybują dany temat lub odbierają wiadomości z tych tematów. Poniższa grafika przedstawia uproszczony układ koncepcyjny klastra Kafka:



W systemach przesyłania wiadomości wiadomości muszą być gdzieś przechowywane. W Kafce przechowujemy wiadomości w Tematach. Każdy temat należy do kategorii, co oznacza, że jeden temat może przechowywać informacje o towarach, a drugi może przechowywać informacje o sprzedaży. Producent, który chce wysłać wiadomość, może wysłać ją do wybranej przez siebie kategorii. Konsument, który chce przeczytać te wiadomości, po prostu zasubskrybuje interesującą go kategorię i zużyje ją. Oto kilka warunków, które musimy znać w zakresie publikowania i subskrybowania architektury:

* Okres przechowywania: wiadomości w temacie muszą być przechowywane przez określony czas, aby zaoszczędzić miejsce niezależnie od przepustowości. Możemy skonfigurować okres przechowywania, który domyślnie wynosi 7 dni od naszego wyboru. Kafka będzie przechowywać wiadomości przez skonfigurowany okres, a następnie je usunie.

* Polityka przechowywania przestrzeni: Możemy również skonfigurować temat Kafka, aby usuwać komunikaty, gdy rozmiar osiągnie próg wymieniony w konfiguracji. Ten scenariusz może jednak wystąpić, jeśli nie wykonałeś wystarczającego planowania pojemności przed wdrożeniem Kafki w swojej organizacji.

* Przesunięcie: Każda wiadomość w Kafce ma przypisany numer zwany przesunięciem. Tematy składają się z wielu partycji; każda partycja przechowuje wiadomości w kolejności, w której przybyły. Konsumenci potwierdzają komunikat z przesunięciem; oznacza to, że wszystkie wiadomości przed tym przesunięciem wiadomości są odbierane przez konsumenta.

* Partycja: każdy temat Kafka składa się z określonej liczby partycji. Musimy skonfigurować liczbę partycji podczas tworzenia tematów. Partycje są dystrybuowane i pomagają osiągnąć wysoką przepustowość.

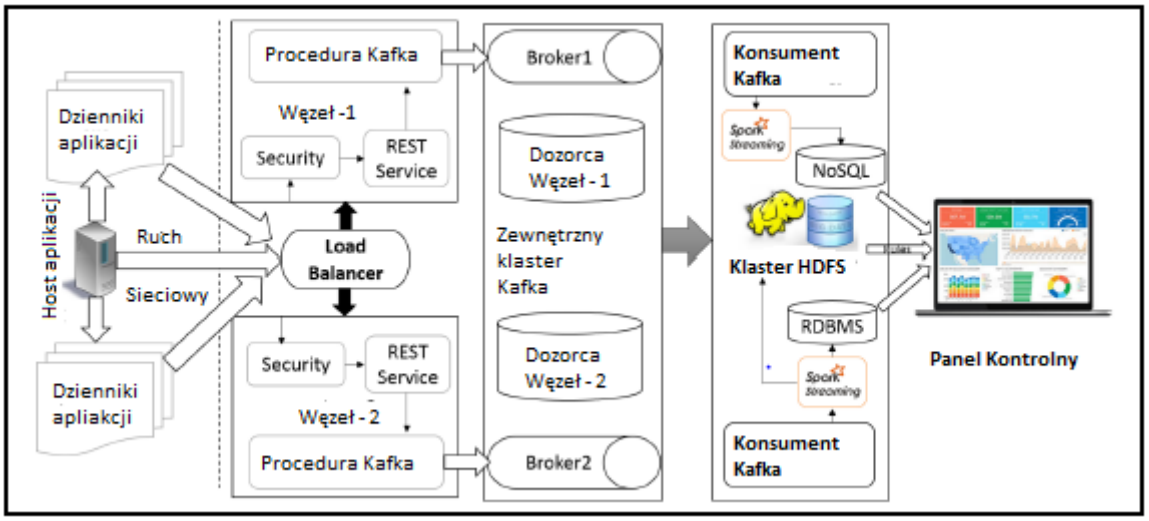
* Zagęszczanie: Zagęszczanie tematów zostało wprowadzone w wersji Kafka 0.8. W Kafce nie ma możliwości przejścia na poprzednie wiadomości, wiadomość zostaje usunięta po upływie okresu przechowywania. Czasami możesz otrzymać nowe wiadomości Kafka z tym samym kluczem, który zawiera kilka zmian, i tylko po stronie konsumenta chcesz przetwarzać najnowsze dane. Kompaktowanie pomaga osiągnąć ten cel poprzez kompaktowanie wszystkich wiadomości za pomocą tego samego klucza i tworzy przesunięcie mapy dla klucza: przesunięcie. Pomaga w usuwaniu duplikatów z dużej liczby wiadomości.

* Lider: partycje są replikowane w klastrach Kafka na podstawie określonego współczynnika replikacji. Każda partycja ma brokera liderów i obserwujących, a wszystkie żądania odczytu i zapisu do partycji będą przekazywane tylko przez lidera. Jeśli lider upadnie, zostanie wybrany inny lider i proces zostanie wznowiony.

* Buforowanie: Kafka buforuje wiadomości zarówno po stronie producenta, jak i konsumenta, aby zwiększyć przepustowość i zredukować IO.

Połączenie Spark Streaming i Kafka tworzy

kompleksowa architektura do wdrażania aplikacji bezpieczeństwa cybernetycznego. Te aplikacje są odporne na uszkodzenia, zapewniają małe opóźnienia i są w stanie obsłużyć dużą liczbę zdarzeń na sekundę. Oto architektura referencyjna dla aplikacji bezpieczeństwa cybernetycznego korzystających z ekosystemu Big Data:



Przyjrzyjmy się teraz niektórym typowym typom ataków cyberbezpieczeństwa i ogólnym strategiom radzenia sobie z nimi.

Typy ataków cyberbezpieczeństwa

„Jednym z głównych zagrożeń cybernetycznych jest myślenie, że ich nie ma. Drugim jest próba wyleczenia wszystkich potencjalnych zagrożeń. (Napraw podstawy, chroń najpierw to, co ważne dla Twojej firmy, i bądź gotowy odpowiednio zareagować na istotne zagrożenia. Pomyśl o danych, ale także o integralności usług biznesowych, świadomości, doświadczeniach klientów, zgodności i reputacji ”. - Stephane Nappo

W miarę cyfryzacji coraz większej liczby systemów i elementów CI rośnie także liczba naruszeń bezpieczeństwa. Atakujący wykorzystują nowatorskie techniki w celu wykorzystania luk w aplikacjach, aby uzyskać dostęp do nieautoryzowanych informacji i uprawnień administracyjnych. W tej sekcji wymienimy niektóre typowe typy ataków i ich ogólne rozwiązania.

Pishing

Jest to jeden z najczęstszych i najskuteczniejszych (z perspektywy atakującego) ataków na aplikacje. W większości przypadków atakujący wysyła do użytkownika wiadomość e-mail lub znajomą komunikację, aby nakłonić go do śledzenia adresu URL i podania kwalifikacji. Chodzi o to, aby użytkownik uwierzył, że wiadomość jest autentyczna. Atakujący czasami tworzy atrapę, ale identyczną stronę internetową,

którą użytkownik zna i nie znajduje powodu, by podejrzewać autentyczność. Gdy użytkownik kliknie adres URL, niektóre złośliwe oprogramowanie jest pobierane na komputer i zaczyna uzyskiwać dostęp do informacji za pośrednictwem połączonych sieci. Atakom tym można zapobiec za pomocą algorytmów uczenia maszynowego. Nagłówki i treść wiadomości e-mail użytkownika mogą służyć jako dane szkoleniowe i mogą szkolić model w zakresie rozumienia typowych wzorców. Ta nauka może pomóc w wykryciu próby wyłudzenia informacji na podstawie trendów behawioralnych w historycznych wiadomościach e-mail.

Ruch lateralny

Gdy atakujący uzyskuje dostęp do sieci przedsiębiorstwa, próbuje wykorzystać luki w danym węźle sieci. Robiąc to, atakujący przenosi się z jednego punktu końcowego sieci do drugiego, uzyskując dostęp do większej liczby usług oraz administracji infrastrukturą sieci i aplikacji. Ten ruch pozostawia ślady w dziennikach sieciowych.

Algorytmy uczenia maszynowego można trenować z ruchami bocznymi do śledzenia danych i wykrywaj podejrzane ruchy użytkownika. Jeśli te ruchy są śledzone przez przesyłanie strumieniowe dzienników sieciowych na żywo przez systemy przetwarzania, włamanie można potencjalnie wykryć w czasie zbliżonym do rzeczywistego.

Ataki Injection

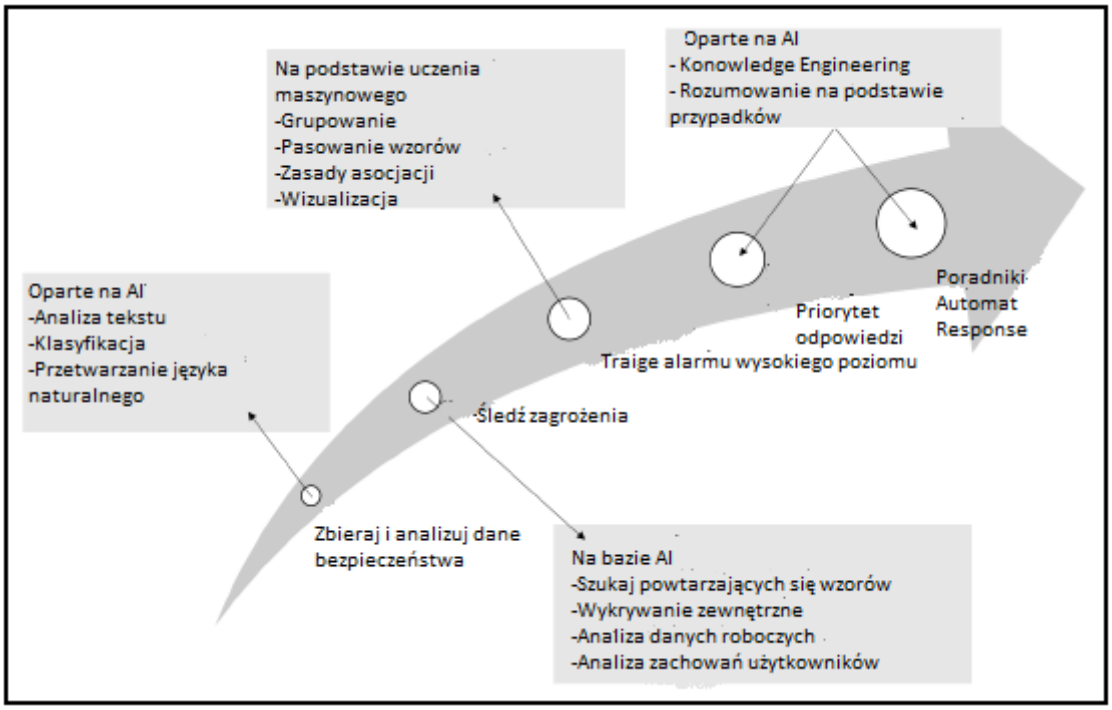
Szkodliwy kod jest dostarczany do aplikacji docelowej za pomocą pól formularza lub innych mechanizmów wejściowych. Wstrzykiwanie SQL jest szczególnym przypadkiem ataku wstrzykiwania, w którym instrukcje SQL są wpychane do systemu za pośrednictwem danych wejściowych pól, a polecenia SQL mogą uzyskać zrzut wrażliwe dane poza siecią. Osoba atakująca może uzyskać dostęp do szczegółów uwierzytelniania, jeśli znajdują się one w bazie danych. Pomimo wszystkich walidacji pól i filtrowania w warstwie serwera WWW, ataki wstrzykiwania są częste i są jednym z wiodących rodzajów ataków. Dzienniki bazy danych mogą być używane do szkolenia modeli uczenia maszynowego na podstawie statystycznych profili użytkowników, które można budować przez pewien okres czasu, gdy użytkownicy wchodzą w interakcję z bazami danych. Nieprawidłowości w schemacie dostępu można wywoływać jako anomalie i generować alarmy. Oprócz iniekcji SQL atakujący czasami uruchamiają skrypty, które podszywają się pod rzeczywistego użytkownika aplikacji i wykonują czynności funkcjonalne w imieniu biznesowym użytkownika. Na przykład, jeśli atakujący może uzyskać dostęp do platformy handlu elektronicznego i rozpoczyna składanie zamówień w imieniu rzeczywistych użytkowników lub wykonuje podobne operacje, takie jak zmiana adresu. W takim przypadku modele uczenia maszynowego muszą zostać przeszkolone, aby nauczyć się indywidualnych zachowań użytkowników, a modele te należy wykorzystać do zidentyfikowania podejrzanych zmian w nawigacji użytkownika i wzorcu działania w aplikacji internetowej.

Obrona oparta na AI

„Dzięki sztucznej inteligencji i uczeniu maszynowemu możemy przeprowadzać wnioskowanie oraz monitorowanie i alarmowanie oparte na wzorcach, ale prawdziwą szansą jest przywrócenie predykcyjne.” - Rob Stroud

Gdy sztuczna inteligencja zostanie zdemokratyzowana, atakujący będą mieli również dostęp do narzędzi i technik wykorzystywania sztucznej inteligencji do atakowania IK. Mechanizm obrony przed takimi atakami również musi się zmodernizować, aby wykorzystać moc danych i obliczeń do szybkiego budowania modeli opartych na sztucznej inteligencji w celu obrony CI i innych aplikacji. Zgodnie z

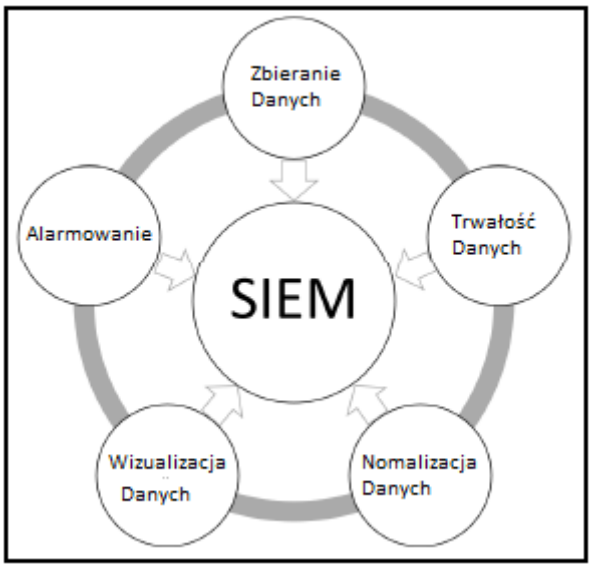
ogólną zasadą poniższy schemat pokazuje etapy mechanizmów obrony opartych na sztucznej inteligencji przed atakami cyberbezpieczeństwa:



Do wykrywania i zapobiegania cyberatakom można wykorzystywać różne algorytmy uczenia maszynowego. Chociaż każda aplikacja różni się konfiguracją sieci i zabezpieczeń, ogólne wytyczne dotyczące zapobiegania atakom cybernetycznym za pomocą różnych algorytmów uczenia maszynowego przedstawiono na poniższym schemacie:

Zrozumienie SIEM

Zarządzanie incydentami i zdarzeniami związanymi z bezpieczeństwem (SIEM) to proces, który pomaga w implementacji bezpieczeństwa cybernetycznego poprzez gromadzenie informacji związanych z bezpieczeństwem (na przykład dzienników sieci i aplikacji) w scentralizowanej lokalizacji lub oznaczanie tych zasobów informacyjnych na krawędzi (lokalizacji, w której dane są generowane w przypadku IoT) i wykorzystuje te informacje do identyfikacji anomalii, które wskazują na naruszenia infrastruktury bezpieczeństwa przedsiębiorstwa. SIEM ułatwia także ciągłe monitorowanie infrastruktury bezpieczeństwa, zapewniając intuicyjne pulpity nawigacyjne wizualizacji. SIEM jako proces jest implementowany jako pakiet oprogramowania zarządzanego przez zabezpieczenia korporacyjne z kontrolą dostępu opartą na rolach. Wspólne cechy charakterystyczne systemu SIEM przedstawiono na poniższym schemacie:



Aplikacja SIEM musi obsługiwać podstawowe elementy składowe w następujący sposób:

* Zbieranie danych: oprogramowanie SIEM powinno obsługiwać różne protokoły komunikacji sieciowej, aby łączyć się z heterogenicznymi systemami w obrębie organizacji. Surowe dane są dostępne w postaci dzienników z aplikacji korporacyjnych, pakietów ruchu sieciowego i kontrolerów sprzętowych. Te surowe dane należy gromadzić w sposób płynny i bezpieczny. Każdy indywidualny system powinien zostać zidentyfikowany i dodany do stosu gromadzenia danych w celu pomyślnego wdrożenia SIEM. Zebrane dane z różnych systemów mogą mieć różne formaty, takie jak tekst, XML, JSON, pliki binarne i tak dalej.

System SIEM musi obsługiwać różnorodne formaty danych.

* Trwałość danych: W zależności od ilości danych oprogramowanie SIEM może wykorzystywać dyski lokalne i sieciowe lub rozproszone systemy plików, takie jak HDFS, do utrwalania danych. Gdy tylko dane z aplikacji i urządzeń są dostępne dla SIEM, musi je przeanalizować w zależności od formatu, zindeksować je i udostępnić do wyszukiwania ad hoc przez użytkownika lub zintegrowaną aplikację. Historyczne i ciągłe dzienniki są ciągłym i stale powiększającym się zasobem, dlatego funkcja indeksowania systemu SIEM musi być zaawansowana i wydajna.

* Normalizacja danych: Jest to jeden z najważniejszych aspektów oprogramowania SIEM. Po uzyskaniu i utrwaleniu danych należy je modelować i znormalizować. Celem normalizacji jest ułatwienie komponentowi wizualizacji wyświetlania krytycznych informacji na desce rozdzielczej. Moduł normalizacyjny może również wykorzystywać zasoby danych do budowy modeli uczenia maszynowego w oparciu o trendy historyczne. Systemy SIEM, które wykorzystują dane do szkolenia uczenia maszynowego modele i analizy predykcyjne będą bardziej poszukiwane w porównaniu z systemami SIEM, które wykonują analizy opisowe i dostarczają alerty oparte na regułach.

* Wizualizacja danych: Wizualizacja jest oknem dla personelu odpowiedzialnego za bezpieczeństwo i zarządzanie przedsiębiorstwem, które mogą wymagać ogólnego widoku ogólnego stanu systemu. Ponieważ decyzje i działania opierają się na tym, co widać na desce rozdzielczej, systemy SIEM muszą wdrożyć przemyślany i dokładny proces definiowania wizualizacji. Ponieważ każde przedsiębiorstwo i przypadek użycia jest wyjątkowy, jedna wizualizacja nie jest odpowiednia dla wszystkich. Narzędzie SIEM musi zapewniać łatwe dostosowania komponentu wizualizacji. Ogólny zestaw elementów wizualizacji przedstawiono na poniższym diagramie:

Atrybuty i funkcje wizualizacji

Atrybuty i funkcje wizualizacji są następujące:

* Pobieranie wartości: oprogramowanie SIEM powinno obsługiwać pobieranie dowolnych wartości atrybutów w zasobach danych. W idealnym scenariuszu oprogramowanie SIEM będzie obsługiwać język zapytań podobny do SQL w celu pobierania danych na podstawie wielu zestawów danych w oparciu o niektóre warunki łączenia.

* Filtrowanie i sortowanie: oprogramowanie SIEM powinno obsługiwać intuicyjne filtrowanie i sortowanie na podstawie jednej lub wielu kluczowych kolumn pożądanych przez użytkownika końcowego.

* Ekstremalne wartości: oprogramowanie SIEM powinno obsługiwać podkreślanie ekstremalnych wartości atrybutów za pomocą kodowania kolorami, aby użytkownik mógł szybko podjąć działania w oparciu o warunki krytyczne.

* Zakres danych: w przypadku kluczowych atrybutów SIEM powinien udostępnić funkcję wyróżnienia wartości zakresu, aby zidentyfikować ewentualne anomalie.

* Dystrybucja danych: oprogramowanie SIEM powinno mieć funkcję pokazującą dystrybucję danych dla kluczowych atrybutów w oparciu o zestaw kryteriów. Może odpowiadać na pytania takie jak: jaka jest dystrybucja różnego rodzaju ataków cyberbezpieczeństwa? Zespół wsparcia może zająć się najważniejszymi przyczynami skutecznego zabezpieczenia IK.

* Reprezentacja anomalii: Anomalie powinny być reprezentowane w taki sposób, aby przyciągały uwagę i dostarczały wystarczających informacji do natychmiastowego ograniczenia ryzyka.

* Klastrowanie i korelacja danych: Dane związane z aplikacjami infrastruktury bezpieczeństwa CI powinny być wizualizowane w klastrach lub grupach powiązanych ze sobą jednostek. Aplikacja powinna obsługiwać niektóre operacje (filtrowanie, sortowanie itd.) W klastrach.

* Alarmowanie: oprogramowanie SIEM powinno obsługiwać mechanizmy generowania alertów o krytycznych zdarzeniach. Użytkownik musi mieć możliwość skonfigurowania progów alertów i skonfigurowania nowych alertów zgodnie z wymaganiami. W przypadku często używanych dzienników, takich jak dzienniki dostępu do serwera WWW, aplikacja powinna mieć predefiniowane alerty, które można szybko skonfigurować, konfigurując wartości progowe. Oprogramowanie powinno także wykorzystywać dane historyczne do szkolenia modeli uczenia maszynowego, które generują alarmy prewencyjne w oparciu o przeszłe trendy.

Ddokonamy przeglądu dwóch pakietów oprogramowania SIEM. Splunk i ArcSight ESM to dwie najpopularniejsze aplikacje SIEM, które są szeroko wdrażane w niektórych elementach CI misji.

Splunk

Splunk to jedno z najpopularniejszych i sprawdzonych rozwiązań SIEM na rynku. Jest zaufany przez ponad 15 000 klientów na całym świecie w zakresie ochrony IK. W tej sekcji omówimy niektóre funkcje obsługiwane przez Splunk do monitorowania bezpieczeństwa i ostrzeganie. Ogólny przegląd platformy Splunk jest przedstawiony w następujący sposób:

Splunk jako platforma zapewnia szereg podproduktów spełniających określone potrzeby organizacyjne. W kontekście tego rozdziału przejrzyjmy funkcje wysokiego poziomu Splunk Enterprise Security i Splunk Light.

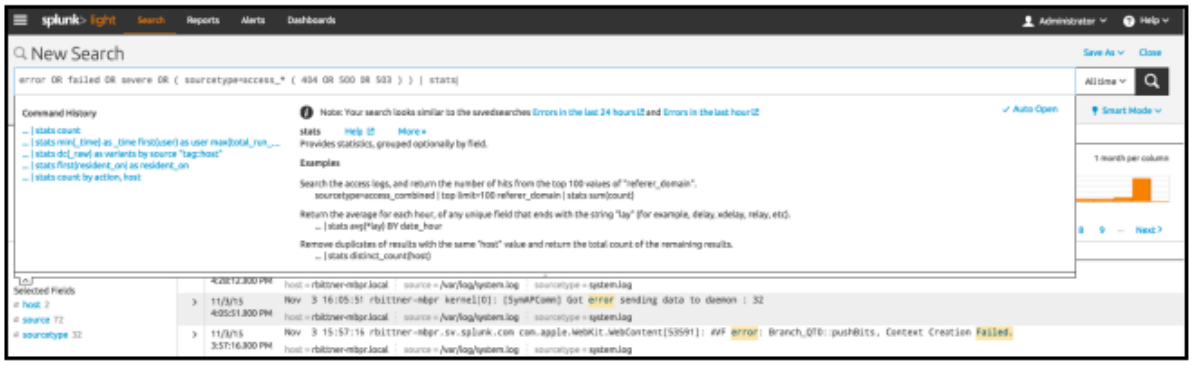
Splunk Enterprise Security

Jest to kompleksowy pakiet, który obejmuje całościowy obraz bezpieczeństwa przedsiębiorstwa, usprawniając operacje bezpieczeństwa przy skróconym czasie działania, udostępniając dane maszynowe do kompleksowej wizualizacji za pomocą interaktywnych pulpitów nawigacyjnych oraz wykorzystując uczenie maszynowe i sztuczną inteligencję do szkolenia modeli predykcyjnych w zakresie bezpieczeństwa prewencyjnego środki.

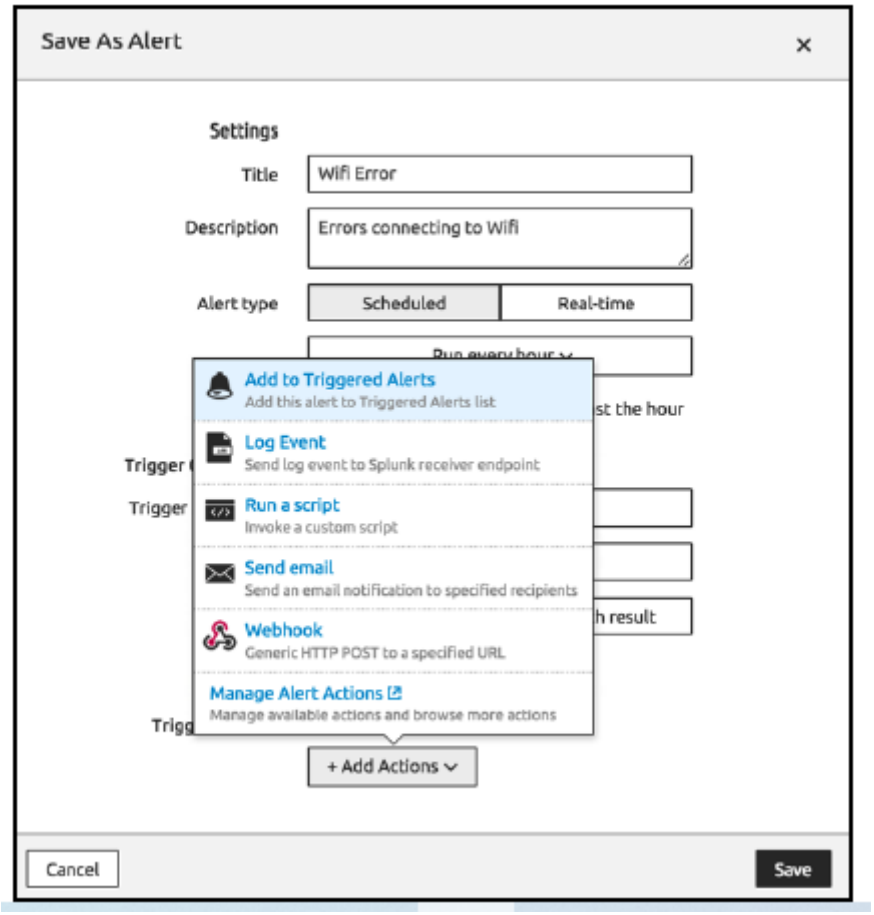
Splunk Light

Splunk Light to specyficzna funkcja produktu, która obsługuje dzienniki dla całego przedsiębiorstwa. Dzienniki zawierają mnóstwo informacji, które można wykorzystać do korekcyjnego i zapobiegawczego cyberbezpieczeństwa. Splunk Light umożliwia przedsiębiorstwom gromadzenie i indeksowanie wszystkich plików dziennika bez względu na ich strukturę i inną semantykę. Warstwa wprowadzania danych jest wystarczająco elastyczna, aby akceptować dzienniki w dowolnym formacie. Istnieje intuicyjny interfejs użytkownika, który odczytuje dzienniki ze skonfigurowanej lokalizacji i prowadzi użytkownika przez różne konfiguracje środowiska wykonawczego, co ułatwia indeksowanie zawartości plików dziennika. Komponent forwardera może zbierać logi z systemów, które nie są bezpośrednio dostępne dla Splunk z powodu ograniczeń sieci. Spedytor może łączyć się ze źródłami zewnętrznymi z wieloma obsługiwanymi protokołami i pobiera dane do Splunk Light w celu wstępnego przetwarzania i indeksowania. Splunk obsługuje schemaless pisze paradygmat struktur Big Data. Schemat jest definiowany w czasie odczytu i może istnieć wiele interpretacji zasobów danych w zależności od kontekstu i przypadku użycia. Inną przydatną funkcją jest obsługa wnioskowania chronologicznego. Splunk może określić sekwencję zdarzeń na podstawie znacznika czasu i komunikatów, w których brakuje znacznika czasu; może również wywnioskować znacznik czasu na podstawie kontekstu.

Wszystkie dzienniki są dostępne w scentralizowanej lokalizacji i można uzyskać do nich dostęp w spójny sposób, niezależnie od źródła i formatu. Dzienniki są stale indeksowane w tle i są dostępne do analizy, filtrowania, sortowania i agregacji. Splunk obsługuje SPL (Splunk Search Processing Language) jako prosty interfejs zapytań podobny do SQL w plikach dziennika. Obsługuje także polecenia analityczne i wizualizacyjne, co ułatwia wykrywanie anomalii w oparciu o różne wzorce i wartości odstające. Wyszukiwanie jest agnostyczne w odniesieniu do wstępnie przetworzonych i indeksowanych dzienników lub dzienników przesyłania strumieniowego. Istnieje wspólny interfejs do przeszukiwania dzienników, który umożliwia zapytania do dzienników w czasie rzeczywistym. Wyniki wyszukiwania można wizualizować za pomocą interaktywnego pulpitu nawigacyjnego. Wizualizacja zapewnia możliwość wycinania i kostkowania od razu po wyjęciu z pudełka i może być łatwo dostosowana do wymagań przedsiębiorstwa. Oto zrzut ekranu wykonania zapytania wyszukiwania w języku przetwarzania:



Aby SIEM był skuteczny, dane zdarzeń z wielu dyskretnych źródeł muszą być dostępne do analizy w scentralizowanym miejscu; Splunk umożliwia korelację złożonych zdarzeń w różnych systemach. Umożliwia to monitorowanie pochodzenia zdarzenia jako takiego pochodzi ze źródła i jego korelacji ze zdarzeniami z innych systemów źródłowych. Ułatwia to natychmiastowe dochodzenie zespołowi ds. Bezpieczeństwa z większą szansą na znalezienie głównej przyczyny anomalii. Splunk Light może automatycznie wykrywać zmiany wzoru bez konieczności interwencji użytkownika. Na przykład określony host aplikacji WWW otrzymuje n żądań w dniu d tygodnia, jeśli nastąpiła znacząca zmiana. Splunk może podkreślić zmianę wzoru, którą można szybko zbadać. Splunk Light umożliwia konfigurację alertów na podstawie typowych wyszukiwań przeprowadzanych przez zespoły administracyjne. Zapytania dotyczące alertów można ustawić tak, aby działały z określoną częstotliwością lub w czasie rzeczywistym, zgodnie z kontekstem przypadku użycia, jak pokazano na poniższym zrzucie ekranu:



ArcSight ESM

ArcSight ESM to produkt HP SIEM, który zapewnia premierowe rozwiązania do zarządzania zdarzeniami bezpieczeństwa. ArcSight analizuje i koreluje każde zdarzenie oraz udostępnia je do wykrywania anomalii. Produkt w znacznym stopniu uzupełnia starania w zakresie zgodności i ryzyka zarządzania. Pomaga zespołom operacyjnym sieci. Najważniejsze cechy ArcSight są następujące:

*Zgodność z przepisami

* Zautomatyzowane zbieranie i archiwizowanie dzienników

*Wykrywanie oszustw

* Wykrywanie zagrożeń w czasie rzeczywistym

* Biznesowy wskaźnik KPI do mapowania i monitorowania zasobów IT

* Analiza wpływu zagrożeń na działalność i automatyczne ustalanie priorytetów

Często Zadawane Pytania

Zróbmy małe podsumowanie.

P: Jakie znaczenie ma Big Data w cyberbezpieczeństwie?

O: Duże zbiory danych i bezpieczeństwo cybernetyczne uzupełniają się nawzajem i odgrywają istotną rolę we wzajemnym znaczeniu i użyteczności. Ponieważ coraz więcej urządzeń łączy się cyfrowo, generuje więcej danych (wolumin); dane generowane przez te podłączone urządzenia muszą być

przetwarzane w neartime (prędkość) i przybiera różne formy, takie jak strukturyzowane, nieustrukturyzowane i półstrukturalne (odmiana). Te trzy V stanowią dane ig w ogóle, co prowadzi do czwartej V. zapobieganie atakom na zasoby komputerowe organizacji.

P: Jakie jest znaczenie infrastruktury krytycznej (CI)? Jakie są kluczowe elementy ochrony CI?

O: Infrastruktura krytyczna to termin używany przez przedsiębiorstwa i agencje rządowe do definiowania aktywów i modeli roboczych, które muszą funkcjonować na optymalnym poziomie, aby zapewnić płynne i harmonijne wrażenia dla interesariuszy, którzy bezpośrednio lub pośrednio korzystają z nich lub mają na nie wpływ systemy. Krajowa sieć energetyczna jest dobrym przykładem IK. Większość systemów CI jest teraz zdigitalizowana, a zatem kontrolowana za pomocą programów komputerowych z minimalnym nadzorem człowieka. Krytyczność tych systemów funkcjonujących wokół sieci zegar czyni je również podatnymi na cyberataki. Systemy chroniące CI przed atakami są również niezwykle ważne z punktu widzenia obrony. Systemy CI generują duże ilości danych dziennika i innych danych operacyjnych. Te dane są najważniejszym zasobem w ochronie IK. Oprócz danych potrzebujemy systemów, które będą w stanie terminowo wykorzystywać i przetwarzać te zasoby danych w celu wykrycia anomalii w zachowaniach systemu i generowania alertów, które wyzwalają ludzkie lub automatyczne działania.

P: Jak wykorzystać uczenie maszynowe i sztuczną inteligencję do skutecznej ochrony CI?

O: Alerty oparte na regułach i systemy monitorowania nie są wystarczające, aby poradzić sobie z atakami cyberbezpieczeństwa i chronić elementy CI. Modele uczenia maszynowego należy szkolić w oparciu o dane historyczne (uczenie nadzorowane), aby przewidzieć wystąpienie złośliwych działań z wyprzedzeniem lub w czasie zbliżonym do rzeczywistego, gdy trwa wtargnięcie. Uczenie maszynowe i sztuczna inteligencja przenosi systemy bezpieczeństwa cybernetycznego na analizę predykcyjną, która pomaga w zapobieganiu atakom.

P: Czy to możliwe, że osoby atakujące wykorzystują sztuczną inteligencję do naruszenia infrastruktury bezpieczeństwa? Jak się przed tym chronić?

O: Tak, atakujący już wykorzystują sztuczną inteligencję i uczenie maszynowe, naruszając infrastrukturę bezpieczeństwa. To wyścig, w którym można lepiej pokonać atakujących i chronić systemy. Dane są zaletą w systemach chroniących CI. Dane pochodzące z heterogenicznych źródeł muszą być wykorzystywane w czasie zbliżonym do rzeczywistego, aby wyprzedzić i chronić IK.

P: Jakie znaczenie ma przetwarzanie strumieniowe w cyberbezpieczeństwie?

O: Zasoby Big Data można przetwarzać w trybie wsadowym i w czasie rzeczywistym. Przetwarzanie w trybie wsadowym jest odpowiednie dla dużych ilości danych i gdy procesy nie są wrażliwe na czas (nie muszą odbywać się w czasie rzeczywistym). Jednak systemy CI stale generują dane jako nieograniczone źródło informacji. Przetwarzanie, przetwarzanie i analiza muszą odbywać się możliwie jak najbliżej zdarzenia, aby mieć szansę na ochronę elementów CI. Przetwarzanie strumieniowe to paradygmat architektoniczny, który zajmuje się nieograniczonymi danymi, które są zużywane jako strumień i przetwarzane, nawet gdy są w ruchu. Jest to przydatne przy wykonywaniu wykrywania anomalii, nawet gdy trwa wtargnięcie, i pomaga zapobiegać potencjalnym atakom na CI.

Podsumowanie

Przeanalizowaliśmy podstawowe koncepcje bezpieczeństwa cybernetycznego i znaczenie Big Data w radzeniu sobie z zagrożeniami dla bezpieczeństwa krytycznych aplikacji. Przetwarzanie dużych danych ma dwa podstawowe typy, przetwarzanie wsadowe i przetwarzanie w czasie rzeczywistym, do

przesyłania danych strumieniowych źródła. Przebadaliśmy podstawowe koncepcje i ramy przetwarzania wsadowego i przetwarzania w czasie rzeczywistym. Przetwarzanie strumieniowe w czasie rzeczywistym jest ważne w przypadku zagrożeń cybernetycznych. Widzieliśmy różne rodzaje typowych zagrożeń bezpieczeństwa i podatności wykorzystywanych przez atakujących. Uczenie maszynowe i sztuczna inteligencja są w dużej mierze zdemokratyzowane i wykorzystywane przez atakujących do wyrafinowanych ataków na IK. To sprawia, że wykorzystanie uczenia maszynowego i sztucznej inteligencji jest krytyczna uwaga przy budowaniu systemów, które zajmują się atakami cyberbezpieczeństwa. Sprawdziliśmy podstawowe elementy składowe systemów SIEM i kilka przykładów, Splunk i ArcSight SEM, jako dwa najpopularniejsze frameworki SIEM. Dziedzina cyberbezpieczeństwa ma pierwszorzędne znaczenie i należy przeprowadzić więcej badań w celu ochrony zasobów danych. Ochrona zasobów danych ma jeszcze większe znaczenie przy coraz większej zależności CI i innych systemów od dostępności dokładnych i wiarygodnych danych. W następnym i ostatnim rozdziale tej książki zajmiemy się obliczeniami kognitywnymi. Inteligencja poznawcza przybliża maszyny do inteligencji ludzkiej, jak to możliwe. To ekscytująca dziedzina badań, w której dokonamy przeglądu niektórych podstawowych koncepcji i dostępnych narzędzi do eksperymentowania i realizacji inteligencji poznawczej w inteligentnych maszynach, które uzupełnią i zwiększą ludzkie możliwości.